README file for:

Harbers, Imke, and Ingram, Matthew C., 2016, "Replication Data for: Geo-Nested Analysis: Mixed-Methods Research with Spatially Dependent Data", doi:10.7910/DVN/HRLHA4, Harvard Dataverse.

Last updated: 2016-12-23

Contents

This readme file contains a description of replication process and materials for the above mentioned paper, and is organized into the following sections:

1. Setup
2. R script
3. Data used for illustration of approach (compressed folder)
4. Supplemental data for mapping results
5. Clarification regarding figures
6. Clarification regarding tables

1. Setup

a. Filepaths

Users are advised to set up a single folder as their working directory, assign the filepath for this working directory to the object “path”, and create five separate sub-folders within this directory: “shapefiles-data”, “output”, “figures”, “maps”, and “stationaritytests”. The R script should be placed in working directory, and then all, uncompressed data files should be placed in sub-folder named “shapefiles-data”. The code makes several directory changes, but with this basic structure in place the code can then be used with few or no adjustments to file paths beyond the initial path settings. If different folder names are used, then filepaths in code should be renamed accordingly.

b. Versions

Using the R command ‘sessionInfo()’, the resulting output provides users with full information on our version of R, the type of machine on which we ran the analysis, and the version of each package used in the analysis. If users attempt replication and a particular package appears not to be working, it may be due to a package update; in such cases, users may want to load version of package identified here. Note that some packages (e.g., Rcpp) are automatically loaded as dependencies of other packages even if they are not manually loaded by user. There should be no need to load these packages manually, but if these packages do not appear in sessionInfo() output or if there is an error related to one of these packages, then users may want to try loading them individually rather than assuming that automatic loading of dependencies worked correctly.

R version 3.3.1 (2016-06-21)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 7 x64 (build 7601) Service Pack 1

attached base packages:
stats
graphics
grDevices
utils
datasets
methods
base

other attached packages:
scales_0.4.0
spgwr_0.6-28
GWmodel_1.2-5
robustbase_0.92-6
plyr_1.8.4
gridExtra_2.2.1
prettyR_2.2
latticeExtra_0.6-28
RColorBrewer_1.1-2
car_2.1-3
weights_0.85
mice_2.25
Rcpp_0.12.6
gdata_2.17.0
Hmisc_3.17-4
Formula_1.2-1
survival_2.39-4
lattice_0.20-33
ggplot2_2.1.0
maptools_0.8-39
spdep_0.6-8
Matrix_1.2-6
sp_1.2-3

2. R script (.R) with replication code

This script is extensively commented in order to assist readers, even those with less familiarity with R.

3. Data used for illustration of approach

The shapefiles and included data used for the illustration in the paper come from a separate study:

Messner, Steven F., Luc Anselin, Darnell F. Hawkins, Glenn Deane, Stewart E. Tolnay, and Robert D. Baller. 2000. *An Atlas of the Spatial Patterning of County-Level Homicide, 1960-1990*. Pittsburgh, PA, National Consortium on Violence Research (NCOVR).

Our paper also makes repeated references to a seminal study of these data: Baller, Robert D., Luc Anselin, Steven F. Messner, Glenn Deane, and Darnell F. Hawkins. 2001. "Structural Covariates of U.S. County Homicide Rates: Incorporating Spatial Effects*." *Criminology* 39: 561–588. doi:10.1111/j.1745-9125.2001.tb00933.x

These data are also publicly available at the website of Luc Anselin's spatial analysis center at the University of Chicago: https://spatial.uchicago.edu/sample-data. Data are also included here for ease of replication, with permission from Luc Anselin and Steve Messner.

Replication data consist of several shapefiles.

Specifically, the main shapefile consists of three separate files:

(1) NAT.SHP
(2) NAT.DBF
(3) NAT.SHX

Note that there is no projection file (.prj) included in this shapefile.

Also, adjusted versions of this shapefile (for south and non-south) are included in the replication materials (named "nat_south2" and "nat_nonsouth2", respectively).

The original data also included a codebook (here, "codebook.pdf") and an html file ("ncovr.html").

4. Supplemental data for mapping results

In addition to the above mentioned shapefile, a separate shapefile captures the boundaries for U.S. states in the maps generated for our paper. Shapefiles with the state boundaries are from the U.S. Census Bureau using 2014 as the reference year. These shapefiles and other geo-spatial data can be accessed at the following URL:

https://www.census.gov/geo/maps-data/data/tiger-line.html

The main state boundary shapefile is "tl_2014_us_state_contig" and is in a zipped folder with same name. Adjusted versions of this shapefile (for south and non-south) are in their own compressed folders. These files are not essential for the replication, but are useful for mapping state boundaries in the visualization of key results.

5. Clarification regarding figures

There are 6 figures in the manuscript: Figures 1, 2, 3.1, 3.2, 4.1, and 4.2. Figures 1 and 2 are straightforward in the replication materials. However, Figures 3.1 – 4.2 require a clarification. There are a wide range of positive and negative values of local coefficients mapped in these figures. These values are easier to discern in colorscale figures, but the manuscript required grayscale figures. To facilitate interpretation, we mapped positive and

negative values separately. Thus, Figures 3.1 and 4.1 map positive values, and Figures 3.2 and 4.2 map negative values. However, the replication code allows users to reproduce both color and grayscale versions of the maps.

The following table clarifies which figures are included in the manuscript. Readers can then distinguish the color figures in the code (by looking closely at code that produces images or simply by looking at file name that includes 'col' (for 'color') rather than 'bw' [for 'black and white']).

Figure	**filename**
3.1	gwrsl_S2SLS-W2X_pos_bw.tif (or .png)
3.2	gwrsl_S2SLS-W2X_neg_bw.tif (or .png)
4.1	gwrse_plain_pos_bw.tif (or .png)
4.2	gwrse_plain_pos_bw.tif (or .png)

6. Clarification regarding tables

There are 4 tables in the manuscript, and there is much more output that could be reported than what is reported in any one of the tables. We present the reduced output for economy of presentation as we illustrate an application of our method using the NCOVR data, but a clarification is in order regarding the larger amount of amount.

For Table 1, the output from four separate models is combined in a single .csv file. The coefficients and standard errors for individual predictors, along with other model diagnostics, can be extracted from the individual model results or from this larger file.

For Table 2, the output generated by the code yields many more rows than those that are reported in Table 2. Specifically, the code reports the 10% of all observations with the largest, statistically significant rho coefficients. Thus, Table 2 reports only the five largest rho coefficients, which is less than 1% of all observations; users may be interested in examining a larger portion of the sample, our code actually produces the top 10% of observations. In any case, the local rho coefficients reported in the last column in Table 2 is the value that appears in the penultimate column in the output, i.e., the column labeled "Wyhat.2_b". The last column in the output reports the local t-statistics so users can confirm that each of the reported local rho coefficients is statistically significant.

For Tables 3 and 4 we followed a similar logic as that used for Table 2. The output generated by the code yields many more rows than those that are reported in Tables 3 and 4 in manuscript. Again, our output yields the top 10% (largest) statistically significant local lambda coefficients (for Table 3) and the bottom 10% (smallest) statistically significant local lambda coefficients (for Table 4). The local lambda coefficients reported in the final column in both Table 3 and Table 4 come from the penultimate column in the output, i.e., the column labeled "Wu_b". The final column in the output reports local t-statistics so that readers can confirm statistical significance.